Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation

Year: 2023 | Citations: 213 | Authors: Hongtao Wu, Ya Jing, Chi-Hou Cheang, Guangzeng Chen, Jiafeng Xu

Abstract

Generative pre-trained models have demonstrated remarkable effectiveness in language and vision domains by learning useful representations. In this paper, we extend the scope of this effectiveness by showing that visual robot manipulation can significantly benefit from large-scale video generative pre-training. We introduce GR-1, a straightforward GPT-style model designed for multi-task language-conditioned visual robot manipulation. GR-1 takes as inputs a language instruction, a sequence of observation images, and a sequence of robot states. It predicts robot actions as well as future images in an end-to-end manner. Thanks to a flexible design, GR-1 can be seamlessly finetuned on robot data after pre-trained on a large-scale video dataset. We perform extensive experiments on the challenging CALVIN benchmark and a real robot. On CALVIN benchmark, our method outperforms state-of-the-art baseline methods and improves the success rate from 88.9% to 94.9%. In the setting of zero-shot unseen scene generalization, GR-1 improves the success rate from 53.3% to 85.4%. In real robot experiments, GR-1 also outperforms baseline methods and shows strong potentials in generalization to unseen scenes and objects. We provide inaugural evidence that a unified GPT-style transformer, augmented with large-scale video generative pre-training, exhibits remarkable generalization to multi-task visual robot manipulation. Project page: https://GR1-Manipulation.github.io